eDTWBI: Effective Imputation Method for Univariate Time Series

Thi-Thu-Hong Phan[1(✉)], Émilie Poisson Caillault[2], and André Bigand[2]

[1] Vietnam National University of Agriculture, Trau Quy, Gia Lam, Hanoi, Vietnam
ptthong@vnua.edu.vn
[2] Univ. Littoral Côte d'Opale, EA 4491-LISIC, 62228 Calais, France
{emilie.caillault,bigand}@univ-littoral.fr

Abstract. Missing data frequently occur in many applied domains and pose serious problems such as loss of efficiency and unreliable results for various approaches. Many real applications require complete data, thus, the filling procedure is a mandatory and precursory pre-processing step. DTWBI is a previously proposed method to estimate missing data in univariate time series with recurrent data. This paper introduces an extension of DTWBI, namely eDTWBI. Firstly, we simultaneously find the two most similar windows to the sub-sequences before and after a gap using DTWBI. Secondly, we impute the gap by average values of the following and previous sub-sequence of the most similar values. Experimental results on three datasets show that our approach outperforms than seven related methods in case of time series having effective information.

Keywords: Imputation · Missing data · Univariate time series · Dynamic time warping · Similarity

1 Introduction

Lots of useful information can be exploited from collected time series and they are used in different domains such as economics [24], finance area [3], health-care [7], meteorology [4,19] and traffic engineering [16]. But the collected data are usually incomplete for various reasons as sensor errors, transmission problems, incorrect measurements, bad weather conditions (outdoor sensors) to manual maintain, etc. Missing data can generate inaccurate data interpretation, biased and unreliable results [8]. Moreover, most of proposed models for time series analysis suffer from one major drawback, which is their inability to process incomplete datasets, despite their powerful techniques. An easy way is to delete or ignore missing data. But this solution comes at high price because of losing valuable information especially for time series where considered values depend on the past ones. So, replacing missing data is a mandatory and precursory pre-processing task. The imputation technique is a conventional method to handle the this problem [11].

H. A. Le Thi et al. (Eds.): ICCSAMA 2019, AISC 1121, pp. 121–132, 2020.
https://doi.org/10.1007/978-3-030-38364-0_11

Imputation methods can be categorized into 2 types: (1) multivariate imputation techniques and (2) univariate imputation approaches. For the first type, these techniques take advantages of relations between variables to estimate missing data [6,7,21,22]. These methods handle incomplete data by filling missing features based on observable ones. They usually train separate models, such as missForest [21], ELM (extreme learning machines) [20], MLP (multi-layer perceptron) [10], etc., for estimating the unobserved attributes.

However, when dealing with missing data in univariate time series, we can only exploit available observations of this variable to predict incompleteness data. Moritz *et* al. pointed out this task is a particularly challenging [13]. And, they performed a review of various methods for univariate time series and showed limitations of some other approaches in [13]. Fewer studies investigates to fill missing data in univariate time series. Simple methods are often used as mean [1], median [5], locf (last observation carried forward), linear interpolation or spline interpolation. For the interpolation methods, missing data are estimated from preceding and succeeding values of the univariate time series. These techniques are effective when the missing data type is isolated (one missing point) or small gap. But when the gap is large, i.e., many consecutive missing values, they do not give good results. For example, if a gap has a sine wave shape, the linear interpolation would complete the gap by a straight line. In addition, we also use statistical methods (e.g. ARMA or ARIMA) to complete missing data in univariate time series but these models require linear data after differencing [2].

Therefore, it is necessary to propose effective imputation methods for univariate time series and consider the characteristics of data, especially for complex distribution data.

In our previous study [15], we proposed DTWBI approach which enables to impute large consecutive missing values in univariate time series. DTWBI is based on the combination of the shape-feature extraction algorithm [14] and Dynamic Time Warping method [17]. In this study, we define a large gap when number of consecutive missing points is larger than the known-process change, so it depends on each application. In order to improve imputation ability, we introduce a novel and effective method for univariate time series, namely eDTWBI which is an extension of DTWBI. Besides, we compare the proposed method with heuristic approach (called Random method) and study the performance of conserving frequency information of all considered methods after the imputation.

This paper is organized as follows. Section 2 focuses on the proposed method. Next, Sect. 3 introduces our experiments, results and discussion. Finally, conclusions are drawn and future work is presented.

2 The Proposed Method: eDTWBI

In the DTWBI algorithm we only envisaged one query either before or after the considered gap. In this study, we modify DTWBI by taking into account two queries, one query before and one query after this gap. Moreover, data before and data after the gap will be treated as two referenced univariate time series.

This would, on the one hand, enrich the learning base and, consequently, increase the prediction ability of the method. On the other hand, this allows to consider dynamics (important key) of data before and after the considered gap to estimate imputation values and to relax temporal constraints between two queries.

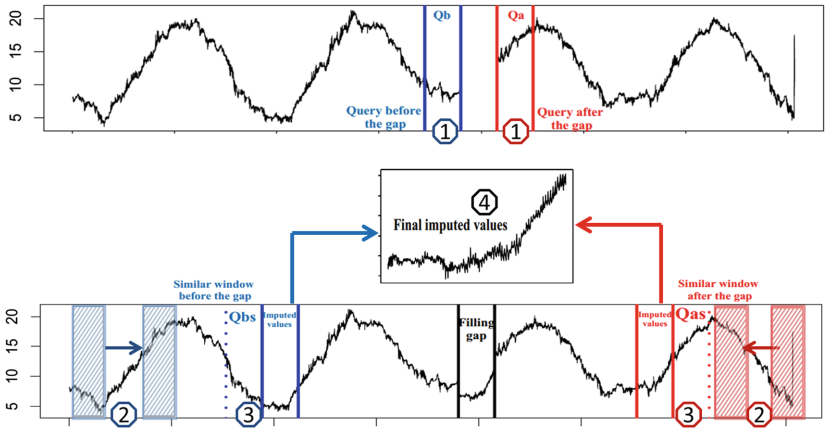

Fig. 1. Scheme of eDTWBI for the imputation task: 1-Queries building, 2-Sliding windows comparison, 3-Most similar windows selection, 4-Gap filling

The eDTWBI algorithm is implemented in order to always ensure accurate results. First, if the position of a gap (with size T) is in the first $2 \times T$ of database, only DTWBI is applied on the remaining data after the gap. If the gap position is in the last $2 \times T$ of series, only DTWBI is performed on data before the gap. In the case of missing data locates in the middle of the series, i.e between $2 \times T$ and $N - 2 \times T$ (where N is the length of the time series), eDTWBI is applied to impute missing data. Figure 1 illustrates the mechanism of eDTWBI to fill large missing data in univariate time series and the detailed algorithm is described in Algorithm 1. This approach consists of three main phases described as follows:

The first phase - Queries building (cf. 1 in Fig. 1): For each T-gap, two referenced time series are extracted from the original signal and two queries are created. The data before the gap (namely Db) and the data after this gap (noted Da) are treated as two separated time series. We noted Qb is the sub-sequence before the gap and Qa is the sub-sequence after the gap, respectively. Qa and Qb queries have the same size T as the gap.

The second phase - Retrieving the most similar windows (cf. 2 & 3 in Fig. 1): For the Da database, sliding reference windows (denoted R) of size T are built. From these R windows, we use DTWBI method [15] to find the most similar window (Qas) to Qa. The same process is carried out to retrieve the most similar window Qbs in Db data.

A key-point of the eDTWBI approach is to envisage the dynamics and shape of data before and after a gap. This means that two queries before and after the

Algorithm 1. eDTWBI algorithm

Input: $X = \{x_1, x_2, \ldots, x_N\}$: incomplete time series
t: index of a gap (position of the first missing of the gap)
T: size of the gap
θ_cos: cosine threshold (≤ 1)
step_threshold: increment for finding a threshold
step_sim_win: increment for finding a similar window
Output: Y - completed (imputed) time series

For each gap ContainsMissing(X) do:
Step 1: Divide X into two separated time series Da, Db: $Da = X[t+T:N], Db = X[1:t-1]$
Step 2: Construct queries Qa, Qb - temporal window after and before the missing data $Qa = Da[1:T]$; $Qb = Db[t-T+1:t-1]$
For Db data do
Step 3: Find the threshold on the Db data
$i \leftarrow 1$; $DTW\_costs \leftarrow NULL$
while $i <= length(Db)$ **do**
    $k \leftarrow i + T - 1$
    Create a reference window: $R(i) = Db[i:k]$
    Calculate global feature of Qb and $R(i)$: $gfQb, gfR$
    Compute cosine coefficient: $cos = cosine(gfQb, gfR)$
    if $cos \geq \theta\_cos$ **then**
        Calculate DTW cost: $cost = DTW\_cost(Qb, R(i))$
        Save the cost to DTW_costs
    $i \leftarrow i + step\_threshold$
$threshold = min\{DTW\_costs\}$
Step 4: Find similar windows on the Db data
$i \leftarrow 1$; $Lopb \leftarrow NULL$
while $i <= length(Db)$ **do**
    $k \leftarrow i + T - 1$
    Create a reference window: $R(i) = Db[i:k]$
    Calculate global feature of Qb and $R(i)$: $gfQb, gfR$
    Compute cosine coefficient: $cos = cosine(gfQb, gfR)$
    if $cos \geq \theta\_cos$ **then**
        Calculate DTW cost: $cost = DTW\_cost(Qb, R(i))$
        if $cost < threshold$ **then**
            Save position of $R(i)$ to $Lopb$
    $i \leftarrow i + step\_sim\_win$
return Qbs - the most similar window to Qb having the minimum DTW cost in the $Lopb$ list.
For Da data do
Perform step 3 and 4 with Da data
return Qas - the most similar window to Qa
Step 5: Replace the missing values at the position t by average vector of the window after the Qbs and the one previous the Qas

studied gap we considered. This allows to detect windows that have the most similar dynamics and shape to the queries.

The third phase - Completing the gap (cf. 4 in Fig. 1): When the two most similar windows are found, we impute the gap by averaging values of the previous window of Qas and the following window of Qbs. In the eDTWBI approach, the average values are used because Schomaker and Heumann indicated that model averaging makes the final results more stable and unbiased [18].

3 Experiments

To illustrate performance of the proposed method, we evaluate it and compare with other imputation methods including DTWBI [15], Kalman [12], na.interp [9], na.locf, na.aggregate and na.spline [25] and heuristic method. To perform the last comparison, we randomly chose 10 windows having the same size of the gap, then compute the average values to fill in the gap.

3.1 Data Description

Four time series are utilized to perform experiments including monthly mean C02 concentrations [23], daily mean air temperature at the Cua Ong meteorological station, monthly mean air temperature and humidity at the Phu Lien meteorological station, in Vietnam. In order to obtain useful information from the datasets and to make the datasets easily exploitable, we analyzed these series. Table 1 summarizes their characteristics. These datasets have a seasonality component (i.e. an annual cycle) without any linear trend. The seasonality component that would be respected after the imputation but they don't have regular amplitude.

Table 1. Characteristics of time series

No	Dataset name	Period	#Samples	Seasonality (Y/N)	Trend (Y/N)	Frequencey
1	CO2 concentrations	1974–1987	160	Y	N	Monthly
2	Phu Lien humidity	1961–2015	692	Y	N	Monthly
3	Phu Lien air temperature	1961–2014	684	Y	N	Monthly
4	Cua Ong air temperature	1973–1999	9859	Y	N	Daily

1. CO2 concentrations - This dataset contains monthly mean CO2 concentrations at the Mauna Loa Observatory from 1974 to 1987 ([23]).
2. Phu Lien humidity - This dataset, containing monthly mean air humidity at the Phu Lien meteorological station in Vietnam, was collected from 1/1961 to 8/2015.
3. Phu Lien air temperature - This dataset is composed of monthly mean air temperature at the Phu Lien meteorological station in Vietnam from 1/1961 to 12/2014.

4. Cua Ong temperature - daily mean air temperature at the Cua Ong meteorological station in Vietnam from 1/1/1973 to 31/12/1999.

3.2 Experiment Process

Actually, assessing the performance of imputation methods can not be done because the real values are missing. Thus, we must generate artificial missing data on complete time series in order to compare the ability of imputation methods. A technique of three steps is used to conduct experiments described in detail as follows:

- *The first step*: Simulated missing data are produced by deleting data segments from each time series with different size of consecutive values.
- *The second step*: All imputation algorithms are applied to estimate the missing values
- *The third step*: The true values and imputed data (generated from different approaches above-mentioned) are compared.

Here, 5 missing data levels are considered on 4 datasets. For CO2 and Phu Lien series, the imputation size ranges from 6%, 7.5%, 10%, 12.5% and 15% of their size respectively. For Ong Ong series, this is a quite big dataset, so gaps are created with size of 3%, 3.75%, 5%, 6.25% and 7.5% dataset length (the largest gap of this time series is 739 missing points i.e. equivalent to more than 2 years of missing data).

For each missing rate in a dataset, 10 missing positions are randomly chosen and all the algorithms are conducted.

3.3 Imputation Performance Indicator

After completing missing data, experiment results are discussed in two parts viz., quantitative performance and visual ability. Specially, the quantitative performance is analyzed in amplitude, shape and frequency criteria. To compare the amplitude between imputation values and actual ones, we use Similarity (Sim), an adapted Normalized Mean Absolute Error (NMAE), Root Mean Square Error (RMSE). Fractional Bias (FB) is applied to compare the shape between prediction data and real data. To assess the ability of frequency conservation of imputation methods, we perform a comparison between the seasonality components of the full series and the imputed signal using NMAE (denoted NMAE(s)). These indicators are computed as following:

1. Similarity - defines the similar percentage between the imputed value (Y) and the respective true values (X). It is calculated by:

$$Sim(Y,X) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{1+\frac{|y_i - x_i|}{max(X)-min(X)}} \tag{1}$$

   Where T is the number of missing values. A higher similarity ($\in [0,1]$) highlights a better ability to complete missing values.

2. NMAE, the Normalized Mean Absolute Error between the imputed value Y and the respective true value time series X is computed as:

$$NMAE(Y,X) = \frac{1}{T}\sum_{i=1}^{T}\frac{|y_i - x_i|}{V_{max} - V_{min}} \tag{2}$$

where V_{max}, V_{min} are the maximum and the minimum value of original time series. A lower NMAE means better performance method for the imputation task.

3. RMSE: The Root Mean Square Error is defined as the average squared difference between the imputed value Y and the respective true value time series X. This indicator is very useful for measuring overall precision or accuracy. In general, the more effective method would have a lower RMSE.

$$RMSE(Y,X) = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i - x_i)^2} \tag{3}$$

4. FB (Fractional Bias) is defined by:

$$FB(Y,X) = 2 * \left|\frac{mean(Y) - mean(X)}{mean(Y) + mean(X)}\right| \tag{4}$$

A model is considered perfect when its FB tends to 0.

3.4 Experiments Results

This part presents experiment results obtained from the proposed approach and compares its ability with the seven published methods.

Tables 2 and 3 show the averaged performance of different imputation methods on 4 datasets for the 5 indicators previously mentioned. These results confirm that eDTWBI is more effective than compared methods in most of the cases, especially in relative high missing rate scenario.

On the CO2 and Phu Lien temperature series, eDTWBI provides the highest Similarity, the lowest NMAE, RMSE and NMAE(s) at nearly every missing ratio (excluding NMAE(s) at 7.5% on the CO2 series). The results demonstrate that the imputation values using eDTWBI method are close to the real values, especially for large missing size (at 12.5% and 15% on Phu Lien temperature and CO2 series). In addition, our method is capable of preserving frequency, which is disclosed on the NMAE(s) index. FB is a quantitative index that allows a shape comparison between predicted and true values. When looking at FB index on Tables 2 and 3, FB values of eDTWBI are the smallest in the majority of missing rates and ranked the second at some levels like 10%, 12.5% on CO2 series, at 15% on Phu Lien temperature, and ranked the 3^{rd} at 7.5%, 10% and 15% on Phu Lien humidity signal. Again, the results show that the improved ability to estimate missing values of eDTWBI method in terms of shape.

Table 2. Average imputation performance indices of various imputation algorithms on CO2 and Phu Lien temperature datasets

Method	Gap size	CO2					Phu Lien temperature				
		Sim	NMAE	RMSE	FB	NMAE (s)	Sim	NMAE	RMSE	FB	NMAE (s)
Random	6%	0.625	0.196	5.22	0.013	0.03	0.779	0.237	4.82	0.018	0.021
Kalman		0.754	0.097	2.768	0.006	0.019	0.58	0.733	14.924	0.48	0.02
DTWBI		0.832	0.055	1.509	0.004	0.009	0.878	0.114	2.576	0.023	0.009
eDTWBI		**0.919**	**0.024**	**0.693**	**0.001**	**0.006**	**0.916**	**0.075**	**1.7**	**0.01**	**0.005**
na.interp		0.731	0.106	2.973	0.006	0.022	0.778	0.244	5.28	0.062	0.02
na.locf		0.721	0.114	3.22	0.006	0.024	0.775	0.257	5.718	0.15	0.019
Aggregate		0.636	0.18	4.802	0.012	0.028	0.791	0.216	4.379	0.016	0.019
na.spline		0.764	0.092	2.66	0.006	0.019	0.599	0.694	14.379	0.433	0.02
Random	7.5%	0.671	0.153	4.042	0.009	0.022	0.798	0.227	4.607	0.014	0.026
Kalman		0.726	0.126	3.607	0.008	0.024	0.534	0.993	19.84	1.273	0.027
DTWBI		0.798	0.068	1.731	0.004	**0.008**	0.883	0.119	2.631	0.022	0.012
eDTWBI		**0.889**	**0.034**	**0.924**	**0.001**	0.01	**0.913**	**0.086**	**1.983**	**0.011**	**0.008**
na.interp		0.737	0.105	3.026	0.005	0.026	0.772	0.281	6.071	0.144	0.026
na.locf		0.725	0.115	3.359	0.008	0.024	0.776	0.273	5.8	0.152	0.025
Aggregate		0.681	0.14	3.846	0.009	0.024	0.797	0.228	4.605	0.013	0.026
na.spline		0.741	0.117	3.414	0.008	0.023	0.547	0.957	19.432	1.241	0.027
Random	10%	0.644	0.196	5.145	0.013	0.041	0.797	0.236	4.802	0.013	0.035
Kalman		0.71	0.15	4.572	0.009	0.033	0.484	1.3	26.479	1.58	0.035
DTWBI		0.735	0.122	3.595	0.009	0.035	0.885	0.12	2.691	0.021	0.016
eDTWBI		**0.804**	**0.082**	**2.271**	0.004	**0.025**	**0.912**	**0.089**	**2.065**	**0.009**	**0.011**
na.interp		0.777	0.09	2.54	**0.002**	0.031	0.787	0.255	5.395	0.029	0.035
na.locf		0.74	0.114	3.202	0.006	0.03	0.775	0.293	6.49	0.189	0.034
Aggregate		0.629	0.204	5.344	0.014	0.038	0.799	0.23	4.644	0.014	0.034
na.spline		0.658	0.591	19.906	0.046	0.043	0.475	1.322	27.19	3.809	0.035
Random	12.5%	0.634	0.199	5.262	0.014	0.047	0.797	0.234	4.756	0.009	0.043
Kalman		0.761	0.107	3.238	**0.003**	0.041	0.622	0.722	15.389	0.372	0.042
DTWBI		0.731	0.122	3.508	0.008	0.036	0.879	0.128	2.835	0.013	0.021
eDTWBI		**0.804**	**0.083**	**2.43**	0.005	**0.031**	**0.901**	**0.101**	**2.304**	**0.008**	**0.016**
na.interp		0.767	0.1	2.886	0.006	0.044	0.78	0.282	6.263	0.171	0.042
na.locf		0.744	0.117	3.338	0.007	0.048	0.763	0.315	6.95	0.229	0.043
Aggregate		0.63	0.206	5.446	0.014	0.043	0.803	0.225	4.547	0.009	0.042
na.spline		0.756	0.113	3.533	0.005	0.041	0.537	1.112	23.34	1.149	0.042
Random	15%	0.674	0.199	5.308	0.014	0.047	0.798	0.234	4.731	**0.007**	0.052
Kalman		0.651	0.233	6.319	0.015	0.053	0.45	1.45	29.078	9.551	0.052
DTWBI		0.747	0.124	3.313	0.008	0.033	0.886	0.12	2.684	0.012	0.023
eDTWBI		**0.831**	**0.082**	**2.297**	**0.005**	**0.031**	**0.897**	**0.107**	**2.388**	0.008	**0.021**
na.interp		0.771	0.115	3.311	0.007	0.05	0.782	0.277	6.192	0.149	0.051
na.locf		0.744	0.135	3.794	0.008	0.05	0.777	0.288	6.404	0.191	0.05
Aggregate		0.699	0.176	4.778	0.011	0.05	0.801	0.227	4.585	0.008	0.05
na.spline		0.662	0.223	6.14	0.015	0.051	0.527	1.097	22.954	1.278	0.051

Cua Ong time series is long so we pay special attention to the shape and dynamics of the imputation values. This is very important when we fill in large missing data. Therefore, we take into account another index, FA2. It represents the fraction of data points that satisfied smoothing amplitude cover. This indicator is calculated as $FA2(Y,X) = \frac{length(0.5 \leq \frac{Y}{X} \leq 2)}{length(X)}$. For the imputation task, if FA2 is closer to 1, the imputation values are closer to the real values. When

Table 3. Average imputation performance indices of various imputation algorithms on Phu Lien humidity and Cua Ong series

Method	Gap size	Phu Lien humidity					Cua Ong temperature				
		Sim	NMAE	RMSE	FB		Sim	NMAE	RMSE	FB	FA2
Random	6%	0.858	0.135	6.8	0.021	0.021	0.83	0.18	54.7	0.042	0.98
Kalman		0.845	0.153	7.4	0.041	0.019	0.83	0.19	58.6	0.152	0.97
DTWBI		0.861	0.132	6.2	0.023	**0.011**	0.901	0.10	33.3	**0.022**	0.993
eDTWBI		**0.877**	**0.114**	**5.6**	**0.018**	**0.011**	**0.906**	**0.09**	**31.3**	0.028	**1.00**
na.interp		0.828	0.176	8.3	0.054	0.02	0.83	0.19	58.6	0.152	0.97
na.locf		0.786	0.236	10.4	0.096	0.021	0.80	0.23	72.2	0.18	0.95
Aggregate		0.865	0.126	6.3	0.019	0.019	0.82	0.19	56.3	0.042	0.98
na.spline		0.534	0.908	37.4	0.4	0.02	0.39	2.43	727.6	2.077	0.21
Random	7.5%	0.84	0.125	5.9	0.02	0.019	0.84	0.18	53.2	**0.014**	0.98
Kalman		0.834	0.132	6.0	0.026	0.02	0.81	0.22	68.5	0.145	0.95
DTWBI		0.851	0.118	5.7	0.016	0.011	0.89	0.10	34.7	**0.016**	0.99
eDTWBI		**0.859**	**0.11**	**5.3**	0.022	**0.01**	**0.912**	**0.09**	**30.2**	0.02	**0.994**
na.interp		0.844	0.125	5.9	0.022	0.02	0.81	0.22	68.5	0.145	0.95
na.locf		0.821	0.149	6.9	0.047	0.019	0.81	0.21	67.4	0.154	0.96
Aggregate		0.84	0.124	5.8	0.02	0.019	0.83	0.18	54.1	**0.014**	0.98
na.spline		0.459	1.193	51.9	0.437	0.02	0.31	3.91	1153.9	2.256	0.15
Random	10%	0.865	0.124	6.1	0.008	0.031	0.83	0.19	57.8	0.052	0.98
Kalman		0.85	0.143	6.9	0.037	0.031	0.83	0.20	62.1	0.056	0.97
DTWBI		0.859	0.134	6.8	0.015	0.025	0.903	0.10	32.6	0.025	0.99
eDTWBI		0.864	0.127	6.3	0.012	**0.023**	**0.912**	**0.09**	**28.7**	**0.022**	**0.999**
na.interp		0.84	0.155	7.3	0.047	0.031	0.83	0.20	62.1	0.056	0.97
na.locf		0.834	0.163	7.6	0.05	0.031	0.83	0.20	64.1	0.134	0.97
aggregate		**0.872**	**0.116**	**5.8**	**0.008**	0.03	0.83	0.18	54.9	0.051	0.98
na.spline		0.423	1.817	74.8	0.69	0.032	0.34	3.42	1005.6	3.035	0.18
Random	12.5%	0.866	0.122	6.0	0.012	0.036	0.82	0.21	61.7	0.031	0.97
Kalman		0.85	0.143	6.8	0.026	0.037	0.82	0.21	66.8	0.153	0.97
DTWBI		0.867	0.122	6.0	0.013	**0.019**	0.9	0.11	35.9	0.023	0.991
eDTWBI		**0.874**	**0.115**	**5.8**	**0.009**	**0.019**	**0.91**	**0.09**	**31.7**	**0.02**	**0.998**
na.interp		0.821	0.179	8.3	0.048	0.037	0.82	0.21	66.8	0.153	0.97
na.locf		0.786	0.221	9.5	0.086	0.035	0.82	0.22	67.9	0.159	0.96
Aggregate		0.873	0.116	5.8	**0.009**	0.036	0.84	0.18	54.3	0.027	0.98
na.spline		0.39	2.103	87.0	1.833	0.04	0.28	5.38	1625.1	2.045	0.13
Random	15%	0.859	0.135	6.2	0.019	0.042	0.84	0.18	53.9	0.013	0.98
Kalman		**0.871**	0.125	6.0	**0.017**	0.039	0.82	0.22	68.2	0.157	0.96
DTWBI		0.863	0.133	6.4	0.026	**0.023**	0.91	0.10	33.1	0.018	0.99
eDTWBI		0.87	**0.123**	**5.9**	0.02	**0.023**	**0.913**	**0.09**	**30.4**	0.018	**0.999**
na.interp		0.865	0.133	6.3	0.02	0.039	0.82	0.22	68.2	0.157	0.96
na.locf		0.854	0.148	7.0	0.041	0.039	0.82	0.21	66.4	0.148	0.97
Aggregate		0.867	0.125	**5.9**	0.018	0.039	0.84	0.18	53.9	**0.01**	0.98
na.spline		0.314	2.674	111.0	1.816	0.04	0.23	6.55	1901.0	6.482	0.08

looking at the results in Table 3, eDTWBI method proves its superior ability compared to other methods for the task of completing missing data on large datasets. It provides the highest Similarity and FA2, the lowest NMAE and RMSE at every missing levels.

Besides, the visualization performance of imputed values generated from different methods is studied. Figure 2 presents the shape of imputation data using 7 different methods on CO2 series. DTWBI well respects the shape of real values.

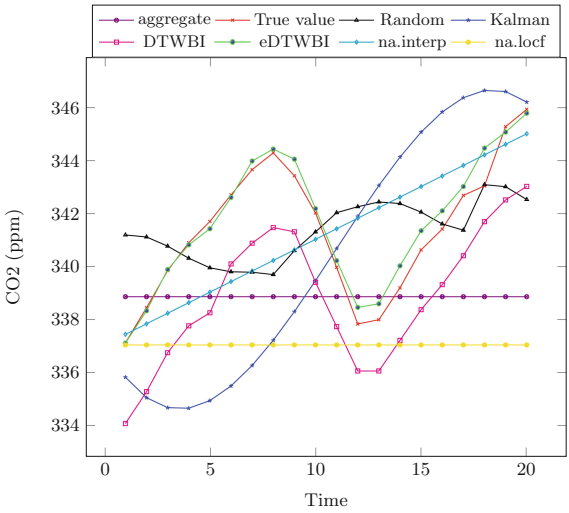

Fig. 2. Visual comparison of imputed values of different univariate methods with true values on CO2 series with the gap size of 20 (12.5%) at postion 89.

But when comparing it with eDTWBI, this approach proves again its ability to deal with missing data. The dynamics of prediction values yielded by eDTWBI is almost identical to the form of true values.

Although the FB value of Kalman method is the 1^{st} on the CO2 and Phu Lien humidity series at 12.5% missing ratio (Tables 2 and 3), but when looking at Fig. 2, it clearly shows that the amplitude and shape of imputation values produced by Kalman differ greatly from the acutal values.

Figure 2 also shows that three methods, including na.aggregate, na.locf and na.interp, always provide a straight line even though quantitative indicators are quite good (Table 2). This means that they do not respect the shape of true values. Random is heuristic method but it provides better results than Kalman or spline in most cases on the Phu Lien series (Table 2).

4 Conclusions and Future Work

This paper proposes a new method, namely eDTWBI, for imputing missing data in univariate time series. The eDTWBI method is an extension of DTWBI by finding the similar values in both databases before and after each gap. It is evaluated and compared with seven other methods on 4 datasets using different criteria (amplitude, frequency and shape constraints). The obtained results clearly demonstrate that our method provides improved performance. However, eDTWBI is based on an assumption of recurring data. In the future, we intend to combine eDTWBI with other algorithms such as interpolation methods to effectually complete missing data in every type of univariate time series.

References

1. Allison, P.D.: Missing Data, Quantitative Applications in the Social Sciences, vol. 136. Sage Publication, Thousand Oaks (2001)
2. Ansley, C.F., Kohn, R.: On the Estimation of ARIMA Models with Missing Values. Springer, New York (1984)
3. Bauer, S., Schlkopf, B., Peters, J.: The arrow of time in multivariate time series, p. 9 (2016)
4. Billinton, R., Chen, H., Ghajar, R.: Time-series models for reliability evaluation of power systems including wind energy. Microelectron. Reliab. **36**(9), 1253–1261 (1996). https://doi.org/10.1016/0026-2714(95)00154-9
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York (2006)
6. Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. J. Stat. Softw. **45**(3), 1–67 (2011)
7. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1) (2018). https://doi.org/10.1038/s41598-018-24271-9, http://www.nature.com/articles/s41598-018-24271-9
8. Hawthorne, G., Hawthorne, G., Elliott, P.: Imputing cross-sectional missing data: comparison of common techniques. Aust. N. Z. J. Psychiatry **39**(7), 583–590 (2005)
9. Hyndman, R., Khandakar, Y.: Automatic time series forecasting: the forecast package for R, used package in 2016. J. Stat. Softw. 1–22 (2008). http://www.jstatsoft.org/article/view/v027i03
10. Jerez, J.M., Molina, I., Garca-Laencina, P.J., Alba, E., Ribelles, N., Martn, M., Franco, L.: Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. **50**(2), 105–115 (2010). https://doi.org/10.1016/j.artmed.2010.05.002, http://www.sciencedirect.com/science/article/pii/S0933365710000679
11. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M.: Methods for imputation of missing values in air quality data sets. Atmos. Environ. **38**(18), 2895–2907 (2004). https://doi.org/10.1016/j.atmosenv.2004.02.026
12. Moritz, S., Bartz-Beielstein, T.: imputeTS: Time series missing value imputation in R. R J. **9**(1), 207–218 (2017). https://journal.r-project.org/archive/2017/RJ-2017-009/index.html
13. Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J.: Comparison of different methods for univariate time series imputation in R. arXiv preprint arXiv:1510.03924 (2015)
14. Phan, T.T.H., Caillault, E.P., Bigand, A.: Comparative study on supervised learning methods for identifying phytoplankton species. In: 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE), pp. 283–288. IEEE, July 2016. https://doi.org/10.1109/CCE.2016.7562650
15. Phan, T.T.H., Caillault, E.P., Lefebvre, A., Bigand, A.: Dynamic time warping-based imputation for univariate time series data. Pattern Recognit. Lett. (2017). https://doi.org/10.1016/j.patrec.2017.08.019
16. Ran, B., Tan, H., Feng, J., Liu, Y., Wang, W.: Traffic speed data imputation method based on tensor completion. Comput. Intell. Neurosci. **2015**, 1–9 (2015). https://doi.org/10.1155/2015/364089
17. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **16**, 43–49 (1978)

18. Schomaker, M., Heumann, C.: Model selection and model averaging after multiple imputation. Comput. Stat. Data Anal. **71**, 758–770 (2014)
19. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting, p. 9 (2015)
20. Sovilj, D., Eirola, E., Miche, Y., Bjrk, K.M., Nian, R., Akusok, A., Lendasse, A.: Extreme learning machine for missing data using multiple imputations. Neurocomputing **174**, 220–231 (2016). https://doi.org/10.1016/j.neucom.2015.03.108, http://www.sciencedirect.com/science/article/pii/S0925231215011182
21. Stekhoven, D.J., Bühlmann, P.: MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012). https://doi.org/10.1093/bioinformatics/btr597
22. Su, Y.S., Gelman, A., Hill, J., Yajima, M., et al.: Multiple imputation with diagnostics (mi) in R: opening windows into the black box. J. Stat. Softw. **45**(2), 1–31 (2011)
23. Thoning, K.W., Tans, P.P., Komhyr, W.D.: Atmospheric carbon dioxide at Mauna Loa observatory. II - analysis of the NOAA GMCC data 1974–1985. J. Geophys. Res.: Atmos. **94**, 8549–8565 (1989)
24. Yang, Y.: Modelling nonlinear vector economic time series, p. 212 (2012)
25. Zeileis, A., Grothendieck, G.: zoo: S3 infrastructure for regular and irregular time series, used package in 2016 (2005). https://doi.org/10.18637/jss.v014.i06, https://www.jstatsoft.org/v014/i06